1.Background, problem definition and objective

# Background

In the realm of consumer lending, the inability to repay loans poses significant challenges for both borrowers and lenders alike. Factors such as job loss, unexpected expenses, economic downturns, and inadequate financial planning can all lead to difficulties in meeting loan obligations

# Problem

Without predictive analysis to anticipate loan repayment challenges, financial institutions risk facing increased default rates, mounting financial losses, and diminished trust from investors and stakeholders due to inadequate risk management strategies.

# Objective

Our aims are to develop advanced predictive models that consider diverse factors such as Age, Sex, Location, and Income Range. By leveraging modern machine learning techniques, the project seeks to provide lenders with actionable insights to mitigate credit risk effectively and make informed lending decisions.

2.Data Collection

# Dataset Description

The dataset contains personal loan financing customer as at 31 December 2023 with defaults and non-defaults loan. The dataset contains 8 columns (features) of 88,233 CBP customers details. The target is identified. The target’s column name is “Default” (Defaults or Non-Defaults).



3.Data preprocessing

For this project, the dataset was collected from CBP has an unbalanced target. The amount of Y (Defaults Loan) is 21,558 and N (Non-Defaults Loan) are 66,675. An imbalanced dataset, where one class significantly outnumbers the others may have several effects on the performance of machine learning. To solve this issue, we using SMOTE function in order to address the class imbalance. The new class distribution after SMOTE will be Y are 46,659 and N are 46,659.

4.Data analysis

The dataset contains both numerical data (Age, Loan rates, Amount Financing, period) and categorical data (Sex, Location, Occupations, Income range). We use univariate analysis and bivariate analysis to analyze our feature.

|  |  |  |  |
| --- | --- | --- | --- |
| Features | All Data | Default Loan | Non-Default Loan |
| Age (Discrete Numerical) | Mean: 42.93  Range: 21 - 71  Std Deviation: 8.05 | Mean:45.57  Range: 23 - 71  Std Deviation:8.25 | Mean:42.07  Range: 21 - 61  Std Deviation:7.80 |
| Loan Rates | Mean:4.74  Range:2.69 – 10.5  Std Deviation:1.12 | Mean:4.4  Range:2.69 -10.5  Std Deviation:0.93 | Mean:4.83  Range:2.69-9.31  Std Deviation:1.16 |
| Amount Financing | Mean:74,177.09  Range:1000 – 400,000  Std Deviation: 56,760.49 | Mean:65,587.13  Range:1000 – 400,000  Std Deviation:51,427.08 | Mean:76,954.48  Range:1000 - 400,000  Std Deviation:58,109.77 |
| Period | Mean: 123.43  Range: 12 - 656  Std Deviation: 35.36 | Mean: 137.54  Range: 12 - 621  Std Deviation:49.72 | Mean: 118.87  Range:12 - 656  Std Deviation:27.75 |

Based on the analysis, we observe those for the numerical feature:

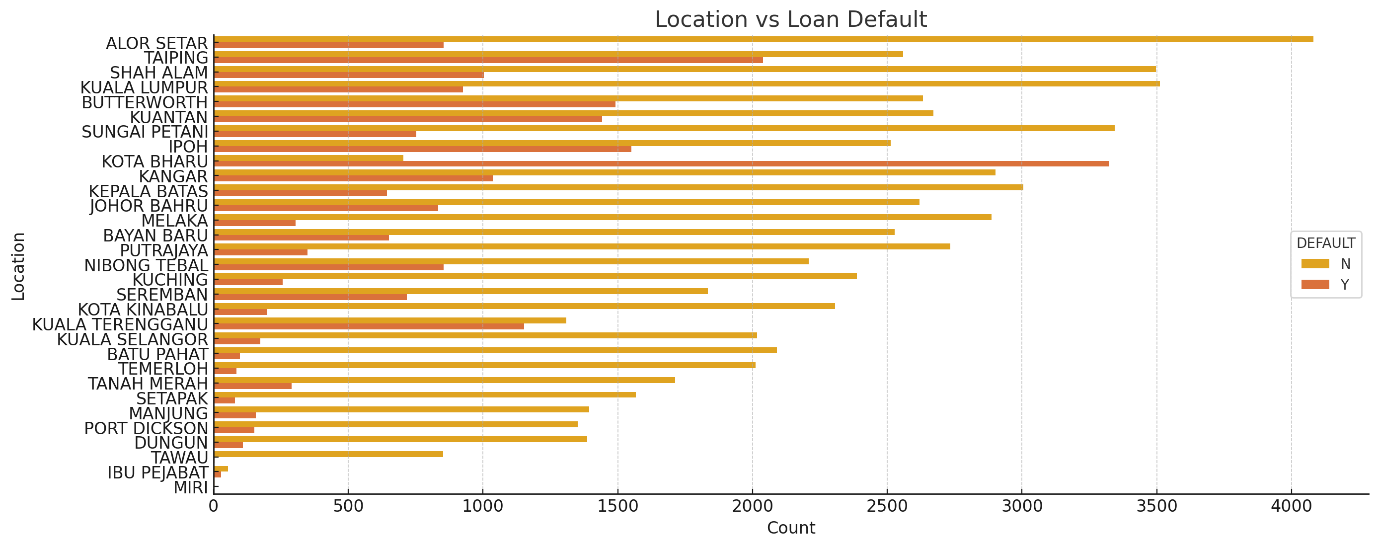
* **Age**: On average, borrowers who default are older (Mean :45.57) compared to those who do not default (Mean:42.07). The age range is also slightly higher for default loans.
* **Loan Rates:** The average loan rate for default loans (Mean: 4.4) is lower than for non-default loans (Mean: 4.83). The standard deviation is also lower for default loans, suggesting less variability in loan rates among defaulters.
* **Amount Financing:** The mean financing amount for default loans (Mean: 65,587.13) is lower than for non-default loans (Mean: 76,954.48). However, the standard deviation is higher for non-default loans, indicating greater variability in financing amounts among non-defaulters.
* **Period:** The average loan period for default loans (Mean: 137.54) is longer compared to non-default loans (Mean: 118.87). The higher standard deviation for default loans suggests more variability in loan periods among defaulters.

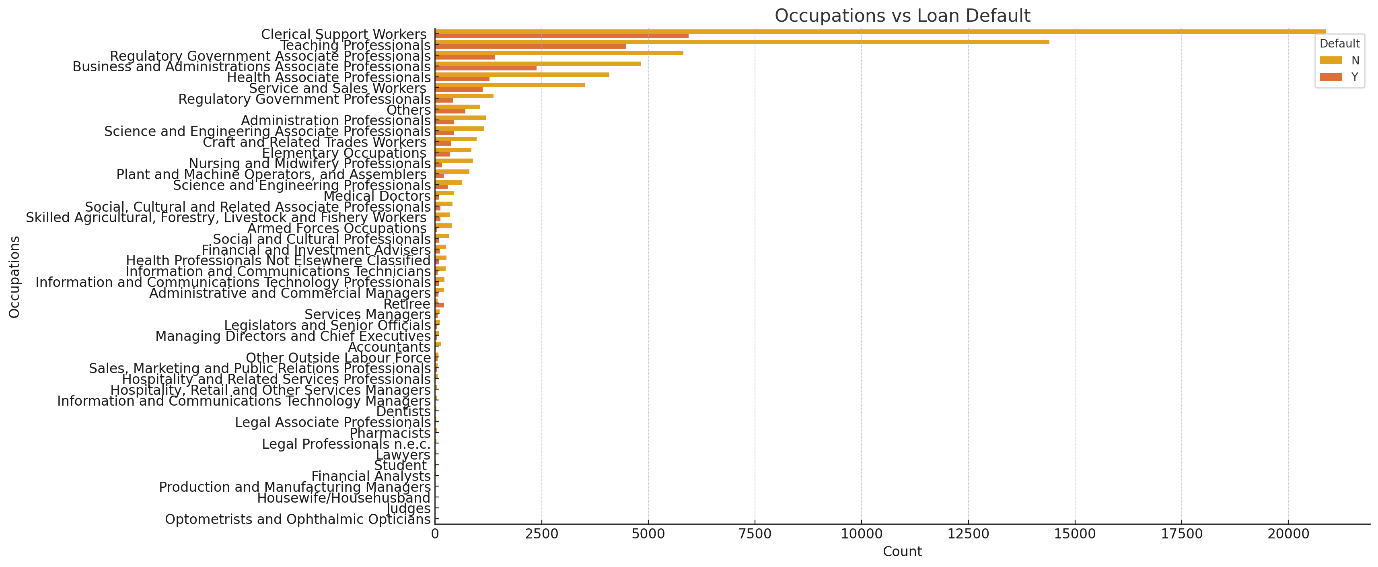
Summary

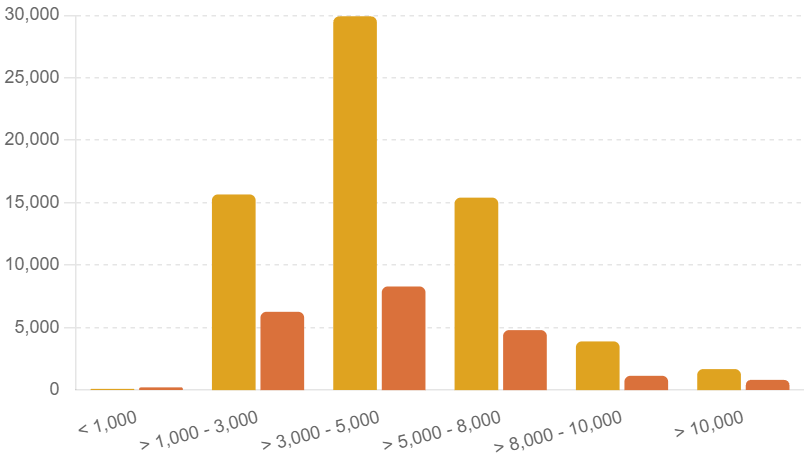
* **Age**: Borrowers who default tend to be older and have longer loan periods.
* **Loan Rates:** Default loans generally have lower loan rates and smaller financing amounts compared to non-default loans.
* **Amount Financing** **and Period:** There is greater variability in the financing amount and loan periods among non-default loans, while the variability in loan rates is higher among non-default loans.

1. Sex

1. Location



1. Occupations
2. Income Range



Based on the analysis, we observe those for the numerical feature:

* **Sex:** The percentage of defaults loan to non-defaults loan of male is slightly higher than the female by 3%
* **Location**: Alor Setar has the highest count of non-defaults loan while Kota Bahru has the highest count of defaults loan.
* **Occupations:** Majority of CBP customers are clerical support workers and teaching professional.
* **Income Range:** Thehighest count for defaults and non-defaults loan are from the income range >RM3,000 – RM5,000.

Summary

* **Sex:** The percentage of defaults loan to non-defaults loan show a significant difference between males and females.
* **Location:** Certain locations have higher counts of defaults compared to others, suggesting location might play a role in the likelihood of default.
* **Occupations:** Some occupations, such as 'Teaching Professionals' and 'Medical Doctors', show varying default rates, indicating occupation may influence default risk.
* **Income Range:** Lower income ranges seem to be associated with higher default rates, while higher income ranges have fewer defaults.

5.Data modelling

### Models Applied and Motivation

#### 1. Decision Tree

**Motivation:** Decision trees are highly intuitive and easy to visualize, making them an excellent choice for understanding the decision-making process of the model. They can handle both numerical and categorical data and are particularly effective in capturing non-linear relationships. Additionally, decision trees require minimal data preprocessing and are robust to outliers, which can be advantageous when dealing with diverse features like AGE, SEX, LOCATION, and OCCUPATIONS in our dataset.

#### 2. Logistic Regression

**Motivation:** Logistic regression is a fundamental and widely-used model for binary classification tasks, such as predicting default status. It provides a straightforward interpretation of the relationship between the features and the target variable through the coefficients. This model is particularly useful for understanding the impact of individual features like LOAN RATES, AMOUNT FINANCING, and PERIOD on the likelihood of default. Additionally, logistic regression is efficient to train and perform well with a relatively small dataset, making it a reliable choice for our prediction project.

#### 3. K-Nearest Neighbors (KNN)

**Motivation:** K-Nearest Neighbors (KNN) is a simple and versatile model that makes predictions based on the similarity of data points. This model is non-parametric, meaning it makes no assumptions about the underlying distribution of the data, which can be beneficial for our diverse dataset. KNN is particularly effective when the data has a clear structure or clustering, as it leverages the distance between data points to make predictions. It is also easy to implement and can provide high accuracy with appropriate feature scaling and selection, making it a valuable addition to our predictive modeling approach.

6.Results

**KNN**

Accuracy: 0.7880

Precision: 0.8002

Recall: 0.7670

F1 Score: 0.7832

Classification Report:

precision recall f1-score support

0 0.78 0.81 0.79 14019

1 0.80 0.77 0.78 13977

accuracy 0.79 27996

macro avg 0.79 0.79 0.79 27996

weighted avg 0.79 0.79 0.79 27996

Confusion Matrix:

[[11342 2677]

[ 3257 10720]

KNN Result:

1)The model correctly predicted the default status 78.80% of the time.

2)Of all the instances predicted as defaults, 80.02% were actually defaults. This indicates the model's effectiveness in minimizing false positives.

3)The model correctly identified 76.70% of the actual defaults. This reflects the model's ability to capture true positive cases.

4) F1 Score: 0.7832

The harmonic mean of precision and recall is 78.32%, providing a balance between precision and recall.

The KNN model achieved an accuracy of 78.80%, with a precision of 80.02% and a recall of 76.70%.

The F1 score of 78.32% indicates a balanced performance between precision and recall.

Confusion Matrix

**True Negatives (TN)** : 11,342

**False Positives (FP)** : 2,677

**False Negatives (FN)** : 3,257

**True Positives (TP)** : 10,720

**Decision Tree**

Accuracy: 0.8531

Precision: 0.8511

Recall: 0.8553

F1 Score: 0.8532

Classification Report:

precision recall f1-score support

0 0.86 0.85 0.85 14019

1 0.85 0.86 0.85 13977

accuracy 0.85 27996

macro avg 0.85 0.85 0.85 27996

weighted avg 0.85 0.85 0.85 27996

Confusion Matrix:

[[11928 2091]

[ 2022 11955]]

**Result**

1)The model correctly predicted the default status 85.31% of the time.

2)Of all the instances predicted as defaults, 85.11% were actually defaults. This indicates the model's effectiveness in minimizing false positives.

3)The model correctly identified 85.53% of the actual defaults. This reflects the model's ability to capture true positive cases.

4)The harmonic mean of precision and recall is 85.32%, providing a balance between precision and recall.

Confusion Matrix

* **True Negatives (TN)**: 11,928
* **False Positives (FP)**: 2,091
* **False Negatives (FN)**: 2,022
* **True Positives (TP)**: 11,955

Logistic Regression

Accuracy: 0.8778

Precision: 0.9400

Recall: 0.8068

F1 Score: 0.8683

Classification Report:

precision recall f1-score support

0 0.83 0.95 0.89 14019

1 0.94 0.81 0.87 13977

accuracy 0.88 27996

macro avg 0.89 0.88 0.88 27996

weighted avg 0.89 0.88 0.88 27996

Confusion Matrix:

[[13299 720]

[ 2700 11277]]

Result

1)The model correctly predicted the default status 87.78% of the time.

2)Of all the instances predicted as defaults, 94.00% were actually defaults. This indicates a low rate of false positives.

3)The model correctly identified 80.68% of the actual defaults. This reflects the model's ability to capture true positive cases.

4)The harmonic mean of precision and recall is 86.83%, providing a balance between precision and recall.

### Confusion Matrix

* **True Negatives (TN)**: 13,299
* **False Positives (FP)**: 720
* **False Negatives (FN)**: 2,700
* **True Positives (TP)**: 11,277

7.Insight or knowledge that has been obtained from the data

KNN

Model **Performance**:

* The KNN model performs reasonably well with an overall accuracy of 78.80%.
* The precision (80.02%) is slightly higher than recall (76.70%), indicating the model is somewhat better at minimizing false positives than at identifying all true positives.
* The F1 score (78.32%) suggests a good balance between precision and recall, but there is room for improvement.

Confusion **Matrix Analysis**:

* The model has 11,342 true negatives and 10,720 true positives, indicating strong prediction capability for both classes.

Decision Tree

Model **Performance**:

* The Decision Tree model performs well with an overall accuracy of 85.31%.
* The precision (85.11%) and recall (85.53%) are balanced, indicating the model is effective at both identifying actual defaults and minimizing false positives.
* The F1 score (85.32%) suggests a good balance between precision and recall, making the model robust.

Confusion **Matrix Analysis**:

* The model has 11,928 true negatives and 11,955 true positives, indicating strong prediction capability for both classes.

Logistic Regression

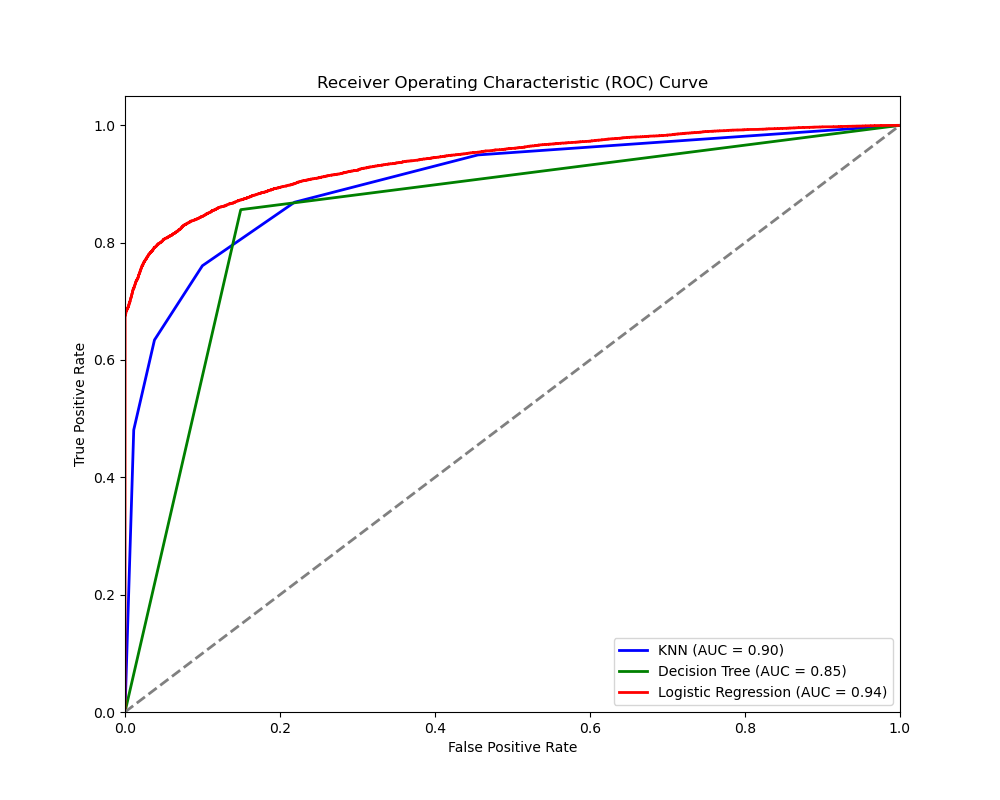
**Model Performance**:

* + The logistic regression model performs well with an overall accuracy of 87.78%.
  + The high precision (94.00%) indicates the model is very effective at minimizing false positives.
  + The recall (80.68%) shows the model is fairly good at identifying true positives but leaves room for improvement in capturing all default cases.

**Confusion Matrix Analysis**:

* + The model has 13,299 true negatives and 11,277 true positives, indicating strong prediction capability for both classes.

Receiver Operating Characteristic Curve:



The ROC curve plot shows the performance of three different models—KNN, Decision Tree, and Logistic Regression—in terms of their ability to distinguish between the positive class (default) and the negative class (will pay). Here are the key insights based on the ROC curve:

1. **Logistic Regression Model**:
   * **AUC = 0.94**: The area under the curve (AUC) for the Logistic Regression model is 0.94, which indicates excellent performance. This means that the Logistic Regression model has a high true positive rate (sensitivity) and a low false positive rate.
   * The ROC curve for Logistic Regression is closest to the top-left corner of the plot, indicating the best overall performance among the three models.
2. **KNN Model**:
   * **AUC = 0.90**: The AUC for the KNN model is 0.90, which is also indicative of good performance. The KNN model performs well, but not as well as the Logistic Regression model.
   * The ROC curve for the KNN model is slightly below the Logistic Regression curve but still demonstrates a strong ability to distinguish between the positive and negative classes.
3. **Decision Tree Model**:
   * **AUC = 0.85**: The AUC for the Decision Tree model is 0.85, which shows that it performs reasonably well but not as well as the Logistic Regression and KNN models.
   * The ROC curve for the Decision Tree model is the lowest among the three, indicating that it has a higher false positive rate compared to the other models for the same true positive rate.

### Summary:

* **Logistic Regression** is the best-performing model with the highest AUC of 0.94, indicating it is the most effective at distinguishing between the positive and negative classes.
* **KNN** also performs well with an AUC of 0.90, making it a strong alternative to Logistic Regression.
* **Decision Tree** has the lowest performance with an AUC of 0.85, suggesting it is less effective at distinguishing between the classes compared to the other two models.